DS-563/CS-543 Lecture 2/2024-01-23

Notes/Reminders
- webpage: https://onak.pl/ds563
  cs543
- HW 0 / HW 1
- electronic device policy

Today:
- CountMin Sketch
- After observing a sequence of element insertions & deletions, we want to provide estimates of frequency for any item

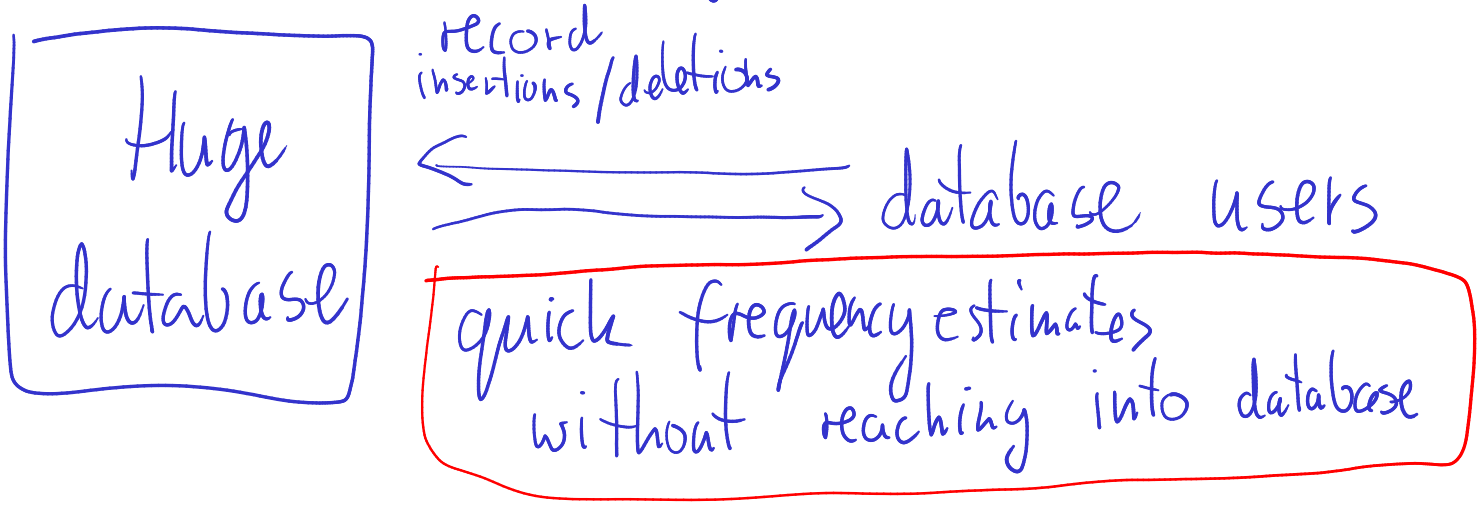
Setting:
- multiset of items (initially empty) from some universe X
- items arrive and depart in arbitrary order

Goal: design a data structure that allows for
- inserting a new item $x \in X$
- removing an item $x \in X$
- querying: "What fraction of items is $y \in X$?"

Example 1: search engine / online store / ...

"What fraction of queries is 'BU mascot'?"

Example 2: tracking statistics inside a large database



Straightforward solution:
explicitly store mapping

$x \in X \rightarrow$ # occurrences of x

(via a dictionary in Python or a map in C++/Java/Rust)

Problem: Lots of space!

Will use less by allowing:

- approximation: return value, say, $\pm 0.01\%$
- randomization: incorrect answer with probability $\delta \in (0,1)$

randomization + approximation ← common themes in this class

First attempt: bucketing

Assume fully random hash function

$$h : X \Rightarrow [k] = \{1, 2, \ldots, k\}$$

Store array $A[1 \ldots k]$ of integers

Initially $A[i] = 0$ for all i

Operations:

Inserting element x: $A[h(x)] \leftarrow A[h(x)] + 1$

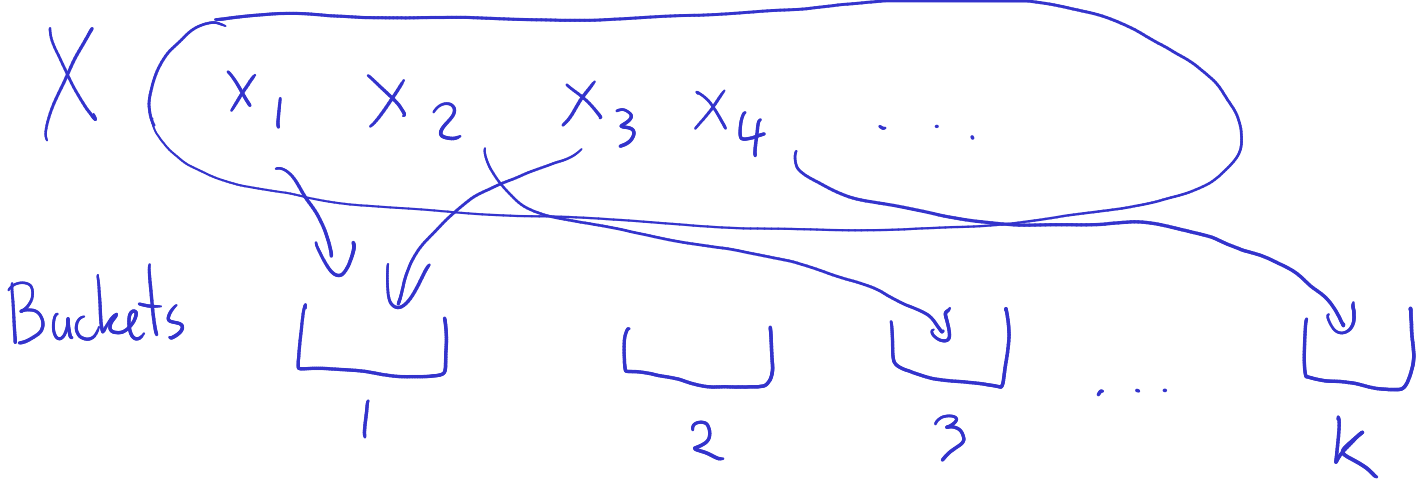
Deleting element x: $A[h(x)] \leftarrow A[h(x)] - 1$

Estimating fraction of y's for $y \in X$:

return $\dfrac{A[h(y)]}{\sum_i A[i]}$

$\sum_i A[i] \overset{\text{def}}{=} S$ = total number of items

estimate $g(y) \overset{\text{def}}{=}$



How good are estimates $g(y)$?

- Can overestimate? Yes. By a lot.
- Can underestimate? No. $g(y) \geq \frac{f(y)}{s}$

$f(y)$ = real number of occurrences of y

Analysis:

$$g(y) = \frac{1}{s} \sum_{\substack{x \in X \\ h(x)=h(y)}} f(x) = \frac{1}{s}\left(f(y) + \sum_{\substack{x \in X \\ x \neq y}} C_{x,y} f(x)\right)$$

collision $C_{x,y} \stackrel{\text{def}}{=} \begin{cases} 0 & \text{if } h(x) \neq h(y) \\ 1 & \text{if } h(x) = h(y) \end{cases}$

random indicator variable

h fully random $\Rightarrow \mathbb{E}[C_{x,y}] = \frac{1}{k}$ for $x \neq y$

$$g(y) = \frac{f(y)}{s} + \underbrace{\frac{1}{s} \sum_{\substack{x \in X \\ x \neq y}} C_{x,y} f(x)}_{\text{error}}$$

error ≥ 0

$$\mathbb{E}[\text{error}] = \frac{\sum_{\substack{x \in X \\ x \neq y}} f(x)\, \mathbb{E}[C_{x,y}]}{s} = \frac{1}{k} \cdot \frac{\sum_{\substack{x \in X \\ x \neq y}} f(x)}{s}$$

$$\leq \frac{1}{k} \cdot \frac{\overbrace{\sum_{x \in X} f(x)}^{= s}}{s} = \frac{1}{k}$$

Markov's inequality

X – non-negative random variable

$a > 0$

$$\Pr[X \geq a] \leq \frac{\mathbb{E}[X]}{a}$$

Hence

$$\Pr\left[\text{error} \geq \frac{2}{k}\right] \leq \frac{1/k}{2/k} = \frac{1}{2}$$

Set $k = \lceil 2/\varepsilon \rceil$ to get additive ε approximation w.p. $\frac{1}{2}$

Next lecture: How to make probability of deviation by more than ε smaller